Validation of FIM-MDS crosswalk conversion algorithm

Ying-Chih Wang, PhD;[1]* Katherine L. Byers, PhD;[2] Craig A. Velozo, PhD, OTR/L[3–4]

[1]*Sensory Motor Performance Program, Rehabilitation Institute of Chicago, Chicago, IL;* [2]*Department of Clinical Administration and Rehabilitation Counseling, Texas Tech University Health Sciences Center, Lubbock, TX;* [3]*Department of Veterans Affairs Health Services Research and Development (R&D) and Rehabilitation R&D, Rehabilitation Outcomes Research Center, North Florida/South Georgia Veterans Health System, Gainesville, FL;* [4]*Department of Occupational Therapy, University of Florida, Gainesville, FL*

Abstract—In this study, we performed a validation analysis of a crosswalk that converts Functional Independence Measure (FIM) scores to Minimum Data Set (MDS) scores and vice versa in order to achieve score compatibility. Data from 2,130 patients were obtained from the Department of Veteran Affairs' Austin Automation Center. The conversion algorithm was tested at the (1) individual patient level, (2) classification level, and (3) facility level. The validity testing resulted in mixed findings. The mean MDS-derived FIM (FIMc) scores were within 1.3 and 0.1 points of the mean actual FIM (FIMa) scores for the motor and cognition scales, respectively. Kappa statistics demonstrated a fair to substantial (0.37–0.66) strength of agreement between functional-related group classifications generated from the FIMa and FIMc scores. Four of the five facilities had an average point difference of 2.4 between the mean FIMa and FIMc scores. While the sample distributions were similar, individual score comparisons fell short of expectations. Only 37% to 67% of the FIMc scores were within 5 points of the FIMa scores. The crosswalk algorithm demonstrated a convenient way to achieve score comparisons across different rehabilitation settings. However, the effectiveness of a single measure or of crosswalk conversions may ultimately depend on the quality of the data.

Key words: Centers for Medicare and Medicaid Services, conversion algorithm, crosswalk, equating, Functional Independence Measure, inpatient rehabilitation facility, Minimum Data Set, patient outcomes, rehabilitation, skilled nursing facilities.

INTRODUCTION

To evaluate and track changes of functional status within the course of rehabilitative healthcare services, researchers and clinicians must integrate information across the continuum of care. While using a single and comprehensive outcome assessment instrument across all facilities would provide the best mechanism for the seamless monitoring of patient outcomes, the current system of site-specific outcome measures makes large-scale reformation extremely difficult.

The Functional Independence Measure (FIM) is currently the most widely used assessment of functional

ability in inpatient rehabilitation facilities (IRFs) and comprehensive rehabilitation hospitals [1]. Through conjoint efforts of the American Congress of Rehabilitation Medicine, the American Academy of Physical Medicine and Rehabilitation, and 11 other national organizations, the FIM was developed as a central core measure of the Uniform Data Set for Medical Rehabilitation to document the functional status (motor and cognition domains) of patient outcomes [2–3]. In 2002, the FIM was incorporated into the IRF-Patient Assessment Instrument, which was implemented within the national prospective payment system (PPS) for inpatient medical rehabilitation reimbursement [4]. Based on age, functional status (provided by FIM), comorbidities, and rehabilitation impairment category, patients are classified into discrete case-mix groups (CMGs) for predicting the resources needed to provide patient care [5].

While the FIM is extensively used in IRFs, the Minimum Data Set (MDS) is completed for all residents in Medicare/Medicaid skilled nursing facilities (SNFs) [6]. In response to the Omnibus Budget Reconciliation Act of 1987, the Health Care Financing Administration (now the Centers for Medicare and Medicaid Services [CMS]) developed a Resident Assessment Instrument (RAI) that all certified SNFs use to assess and plan for the care of residents. As a central assessment core in the RAI, the MDS documents a variety of residents' health information for care planning [7]. By 1998, Medicare/Medicaid SNFs were required to encode and transmit MDS data to the state. After the PPS was mandated in IRFs, information from the MDS assessment (including an activity of daily living [ADL] score, a depression index, and a cognition performance score) was used to classify residents in SNFs into resource utilization groups for better care planning [8–9].

As a result, the FIM is widely used in IRFs and the MDS is mandated in SNFs. When patients are transferred between these facilities, functional status information is discontinuous. Longitudinally, we are unable to track and monitor trajectories of functional outcomes regarding individual patients. In addition, directly comparing functional outcomes across these rehabilitation settings is difficult.

In 2000, CMS recommended that IRFs use the MDS for Post Acute Care (MDS-PAC) as a uniform instrument to integrate healthcare information across postacute care [10]. Structurally similar to the original MDS, the content of the MDS-PAC was expanded to cover the care needs and outcomes across postacute settings [11]. For instance, the functional status section comprises not only basic ADLs (e.g., eating, transferring, dressing, and walking) but also instrumental ADLs (e.g., meal preparation, financial management, and car transfer) [10]. Nonetheless, because of considerable resistance, CMS is no longer considering the use of the MDS-PAC to monitor outcomes for postacute care [5,12].

Although the concept of developing a "universal" postacute assessment tool is conceptually sound, practical issues arise [13]. Both the FIM and MDS have undergone development for more than a decade and are presently the basis for resource allocation in IRFs and SNFs. For example, the FIM has been embedded into algorithms for computing medical payments. Hence, attempts to change or replace these measures would likely face considerable resistance from healthcare practitioners who use these instruments on a daily basis and administrators who base their administrative structure on these instruments (e.g., software, risk-adjustment algorithms, training, and data collection). Implementing a new assessment tool requires not only large-scale staff retraining and transformation in administrative procedures but also consideration of the consequences and impact of such a tool on rehabilitation services and reimbursement algorithms. Adopting a new uniform measure across all postacute care is likely to cost millions, if not billions, of dollars.

Developing a universal assessment tool also requires an adequate item set that covers the entire spectrum of postacute care settings with a variety of patients at different disability levels. This requirement leads to the issue of administrative burden (large item banks) and the adoption of items that may be inappropriate for particular IRF settings. For example, "easy" mobility items, such as "rolling in bed" might be essential to the assessment of individuals with severe disabilities in SNFs but will provide little relevant information for most of the patients treated in outpatient rehabilitation clinics. More challenging items, such as "walking outdoors for 200 meters," may provide useful information for evaluating patients who are 3 months postdischarge (i.e., follow-up evaluations) but would not be appropriate for evaluating patients who are in IRFs.

Furthermore, a new universal assessment tool would require considerable research to support its reliability and validity. Without rigorous psychometric research studies, health service researchers are unlikely to support such an instrument. All of these are practical issues when the replacement of existing instruments with a universal assessment tool across all postacute care settings is being considered.

Another option that avoids many of the challenges and complications of a universal assessment tool is the creation of crosswalks or statistical links between existing

1067

WANG et al. Validation of FIM-MDS crosswalk

functional status instruments. In 2006, Velozo and colleagues developed a FIM-MDS crosswalk conversion algorithm to link the FIM and the MDS [14]. Their intent was to develop an effective, efficient method of tracking and evaluating veterans' functional status changes across IRFs and SNFs. The Rasch true-score equating method [15] was used to develop the FIM-MDS crosswalk, with the basis of the crosswalk being the similar ADL content of the FIM and the MDS. A sample of Department of Veteran Affairs (VA) patients who had been assessed on both instruments within 5 days was used to generate the crosswalk. A FIM-MDS raw score crosswalk conversion table was created whereby total FIM raw scores could be translated into total MDS raw scores and vice versa.

A critical step in determining the feasibility of this crosswalk methodology for practical application is to test its validity. Therefore, we performed a follow-up validation analysis based on the FIM-MDS crosswalk conversion algorithm developed by Velozo and colleagues [14]. The validity of the conversion algorithm was tested at three different levels: (1) individual patient, (2) classification, and (3) facility.

METHODS

Data Preparation

This study was a secondary retrospective data analysis using FIM and MDS data from the Functional Status Outcomes Database and the MDS database collected by VA's Austin Automation Center between June 1, 2002, and December 31, 2004. The two data sets were merged based on patients' scrambled Social Security numbers. Ethical consent for this study was obtained from the University of Florida Institutional Review Board and the VA Subcommittee of Human Studies. Access to VA MDS data was approved by the VA Veterans Health Administration.

We selected patient records that had both FIM and MDS assessments with ≤5 days between administrations. If a patient had multiple records, we selected the FIM-MDS record with the fewest number of days between administrations. Three major impairment groups in the database were selected for analysis: stroke, amputation, and orthopedic impairment. Data were cleaned of obviously invalid data, such as an age of 159 years or functional status items with negative coding. Only records with complete data for all FIM and MDS functional status items were included in the analysis. On the basis of these criteria, we analyzed a final data set of 2,130 patients out of the original 151,770 available records.

Data were further split into two data sets. Data from June 1, 2002, to May 31, 2003 (phase I) were used to generate the FIM-MDS crosswalk tables (motor and cognition) by following the procedure described by Velozo et al. [14]. Data from June 1, 2003, to December 31, 2004 (phase II) were used to perform the validity testing.

Analytical Procedure

Before using the conversion table, we implemented several procedural steps. First, the MDS rating of "8" (activity did not occur) was recoded to a "4" (total dependence). This recoding was based on the rationale that the most likely explanation for a basic ADL not being observed during the entire observation period is that the patient was unable to perform the task. This assumption has been made in several previous MDS-related studies [16–17]. Next, four "recall" cognition items on the MDS instrument had to be recoded so that all cognition MDS items had the same rating-scale direction, with a lower score indicating better performance. Lastly, item ratings from each motor and cognition scale were summed to obtain a total motor and cognition raw score. The total score of the FIM subscales consisted of 13 motor items and 5 cognition items, whereas the total score of the MDS subscales consisted of 13 motor items and 16 cognition items.

The conversion algorithm was tested at three different levels: (1) individual patient, (2) classification, and (3) facility. At the individual patient level, we calculated the absolute value of point differences between the actual FIM (FIMa) scores and the MDS-derived FIM (FIMc) scores (|FIMa – FIMc|) and the percentage of FIMa-FIMc scores within 5 and 10 points. To compare whether the score distributions were similar between the actual and converted scores, we used paired t-tests to test the equivalence of the score distributions.

At the classification level, the functional-related group (FRG) classification system [18] was used to examine whether the FIMc would classify the same patient into a classification level similar to that derived from the FIMa. The FRG is a patient classification system being used by the CMS and is the basis for computing the PPS payment in IRFs [19]. Using FRGs, patients were first classified into 1 of 20 impairment categories and further divided by the two FIM subscales (motor and cognition) and by age at admission to the IRF. Patients assigned to different FRGs are expected to have different rehabilitation outcomes and total costs of care. Therefore, whether the converted scores could classify the individuals into the same FRG as the

actual scores provides a pragmatic test of the precision of the conversion algorithm. After classifying patients into one of the FRGs, we calculated the percentage of individuals being classified into the same FRG category, one category apart (±1 level), and two categories apart (±2 levels). Chi-square statistics were used to test whether any association existed between classification results based on the actual and converted scores. Meanwhile, kappa statistics were used to quantify the strength of association. Landis and Koch provide guidelines for interpreting the strength of agreement: a kappa statistic ranging from 0.21 to 0.40 demonstrates a fair strength of agreement, from 0.41 to 0.60 indicates a moderate strength of agreement, and from 0.61 to 0.80 indicates a substantial strength of agreement [20].

Finally, the conversion algorithm was evaluated at the facility level. After the actual MDS scores were converted to FIMc scores via the conversion algorithm, the score distributions of the FIMc were compared with the actual score distributions of the FIMa by facility. Paired t-tests were used to test the equivalence of the facility-level mean scores obtained from the FIMa and FIMc scores. Since the final data set resulted in many facilities with small sample sizes, only the five facilities that had sample sizes of more than 50 patients were used for this analysis.

RESULTS

Sample

The first data set (phase I) used to generate the FIM-MDS crosswalk tables (motor and cognition) contained 654 subjects: 302 with stroke, 113 with amputation, and 239 with orthopedic impairment. The age of the sample ranged from 20.0 to 100.0 years, with a mean age of 68.0 ± 12.0 years (all data presented as mean ± standard deviation unless otherwise noted). The mean assessment date difference between the FIM and the MDS was 2.9 ± 1.7 days. Ninety-six percent of the sample was male and seventy-four percent was white.

The second data set (phase II) used to perform the validity test contained a sample of 1,476 subjects: 804 with stroke, 268 with amputation, and 404 with orthopedic impairment. The age of the sample ranged from 26.0 to 97.0 years, with a mean age of 70.2 ± 11.7 years. The mean assessment date difference between the FIM and the MDS was 3.2 ± 1.7 days. Ninety-seven percent of the sample was male and sixty-nine percent was white. **Table 1** presents the baseline demographic information of these two samples.

Table 1.

Baseline demographic characteristics of subjects used to generate FIM-MDS crosswalk tables (phase I, $N = 654$ subjects) and to perform validity test (phase II, $N = 1,476$ subjects).

Characteristic	Phase I		Phase II	
	n	%	*n*	%
Sex				
Male	630	96.6	1,434	97.2
Female	22	3.4	40	2.7
Race				
White	485	74.2	1,019	69.0
Black	120	18.3	313	21.2
Hispanic	26	4.0	56	3.8
Native American	8	1.2	8	0.5
Asian	2	0.3	2	0.1
Other	5	0.8	40	2.7
Missing	8	1.2	38	2.6
Impairment Group				
Stroke				
Left-Body Involvement	140	21.4	343	23.2
Right-Body Involvement	134	20.5	285	19.3
Bilateral Involvement	7	1.1	32	2.2
No Paresis	8	1.2	80	5.4
Other Stroke	13	2.0	64	4.3
Amputation (Lower-Limb)				
Unilateral Below Knee	71	10.9	140	9.5
Unilateral Above Knee	30	4.6	73	4.9
Bilateral Below Knee	9	1.4	16	1.1
Bilateral Above/Below Knee	2	0.3	9	0.6
Bilateral Above Knee	1	0.2	9	0.6
Orthopedic				
Unilateral Knee Replacement	84	12.8	166	11.2
Unilateral Hip Replacement	74	11.3	106	7.2
Unilateral Hip Fracture	31	4.7	56	3.8
Major Multiple Fractures	6	0.9	12	0.8
Femur Fracture	5	0.8	4	0.3
Pelvic Fracture	3	0.5	3	0.2
Bilateral Knee Replacement	2	0.3	4	0.3
Bilateral Hip Fracture	1	0.2	3	0.2
Other Orthopedic	33	5.0	46	3.1

FIM = Functional Independence Measure, MDS = Minimum Data Set.

Validation at Individual Level

Score Distribution

Figure 1(a) and **(b)** shows the actual MDS score distributions for the motor and cognition scales, respectively. The MDS motor domain had a score range of 0 to 52, with a score of 0 indicating that a person can perform every task

1069

WANG et al. Validation of FIM-MDS crosswalk
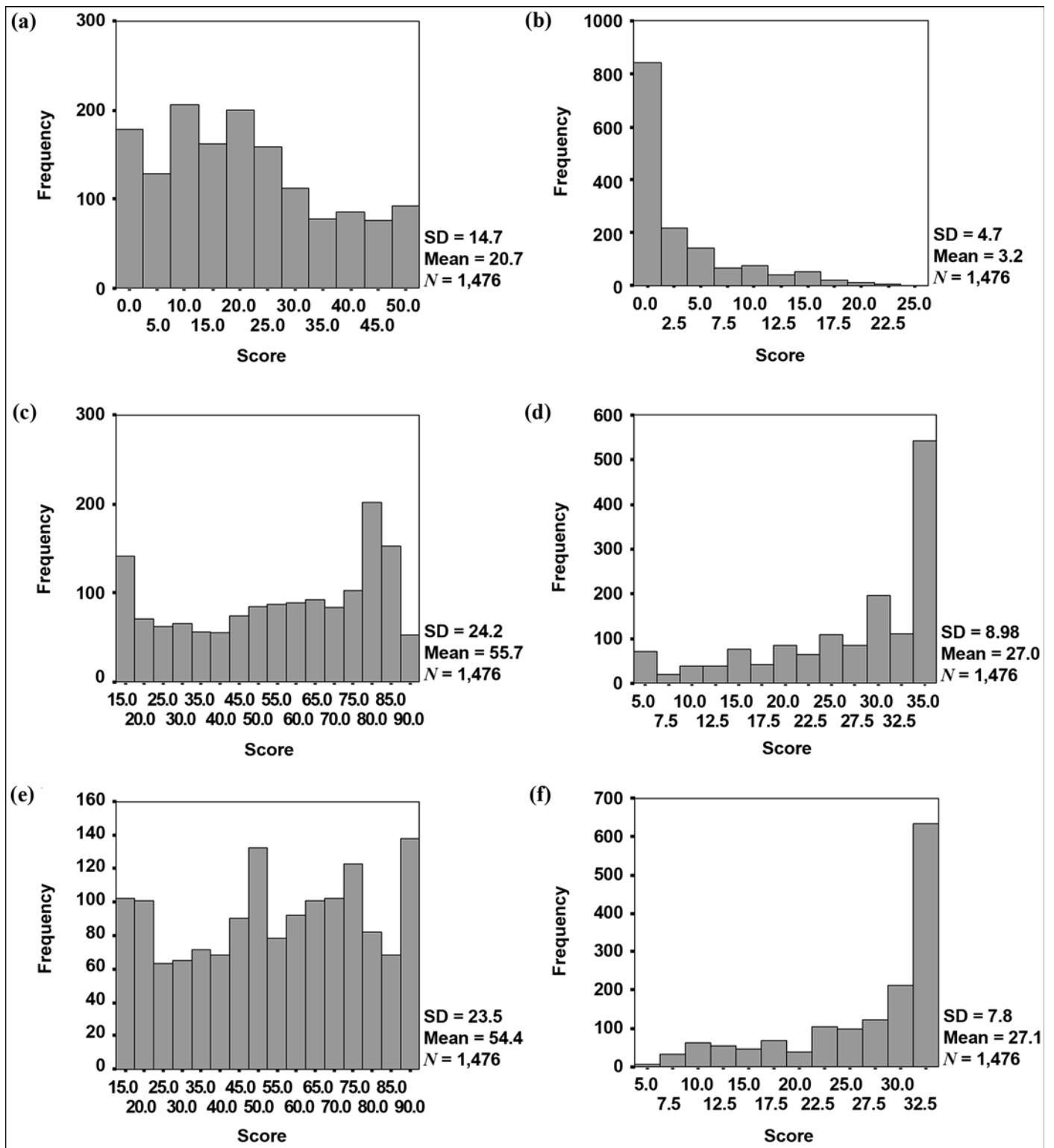
Figure 1.
Score distribution of Functional Independence Measure (FIM) and Minimum Data Set (MDS). **(a)** Actual MDS motor score distribution. **(b)** Actual MDS cognition score distribution. **(c)** Actual FIM motor score distribution. **(d)** Actual FIM cognition score distribution. **(e)** MDS-derived FIM motor score distribution. **(f)** MDS-derived FIM cognition score distribution. SD = standard deviation.

independently without any assistance. With a mean of 20.7, the motor domain had a slightly skewed distribution toward higher functioning individuals (skewness = 0.43). The MDS cognition domain (ranging from 0–25), however, showed a severe ceiling effect (skewness = 1.84). Approximately 42.9 percent of subjects (633/1,476) scored intact on every cognition item, and 79.0 percent of subjects (1,166/1,476) were within 5 points of the highest MDS cognition score (i.e., 0–5).

Figure 1(c) and (d) shows the FIMa score distributions for the motor and cognition scales, respectively. The FIM motor domain had a score range of 13 to 91, with a score of 91 indicating that an individual can perform every task independently. Similar to the MDS score distributions, the FIM motor scale had a slightly skewed distribution toward higher functioning individuals (skewness = –0.37). The FIM cognition scale had a score range of 5 to 35, with a score of 35 indicating that an individual is rated intact on every cognition item. Again, like the MDS cognition scale, the FIM cognition scale had a severe ceiling effect (skewness = –1.03). Approximately 29.6 percent of subjects (437/1,476) obtained the maximum FIM cognition score of 35, and 53.6 percent of the subjects (791/1,476) were within 5 points of the maximum score (i.e., 30–35).

Results from the Kolmogorov-Smirnov test of normality indicated that both FIM and MDS score distributions deviated from a normal distribution ($p < 0.001$). Additionally, the FIMa motor and cognition scales correlated with the actual MDS motor and cognition scales at –0.80 and –0.66, respectively.

Point Different Between Actual and Converted Functional Independence Measure Scores

The MDS scores were converted into FIM compatible scores via the crosswalk conversion algorithm. Figure 1(e) and (f) shows the FIMc for the motor and cognition scales, respectively. Compared with the FIMa score distributions in Figure 1(c) and (d), the FIMc scores showed a similar score distribution. The FIMc motor scale had a mean of 54.4, which was close to the FIMa motor score distribution with a mean of 55.7. While the mean difference of the distributions was only 1.3, the results from the Wilcoxon signed rank test showed a significant difference between these two score distributions ($z = –4.11$, $p < 0.001$). Pearson correlation coefficient was 0.79 between the FIMa and FIMc motor scores. Approximately 33.7 percent of the actual and converted scores showed a 5-point difference or less, and 56.9 percent showed a 10-point difference or less (Figure 2). The mean point difference was 11.6 ± 10.4 points.

The FIMc cognition scale had a mean of 27.1 ± 7.8, which was virtually identical to the FIMa cognition score distribution with a mean of 27.0 ± 9.0. The FIMc cognition scores correlated moderately with the FIMa cognition scores (Pearson correlation coefficient $r = 0.67$). Approximately 67.1 percent of the actual and converted scores showed a 5-point difference or less, and 87.7 percent showed a 10-point difference or less (Figure 2). The mean point difference was 4.9 ± 4.8 points. Table 2 summarizes the validation results at the individual level.
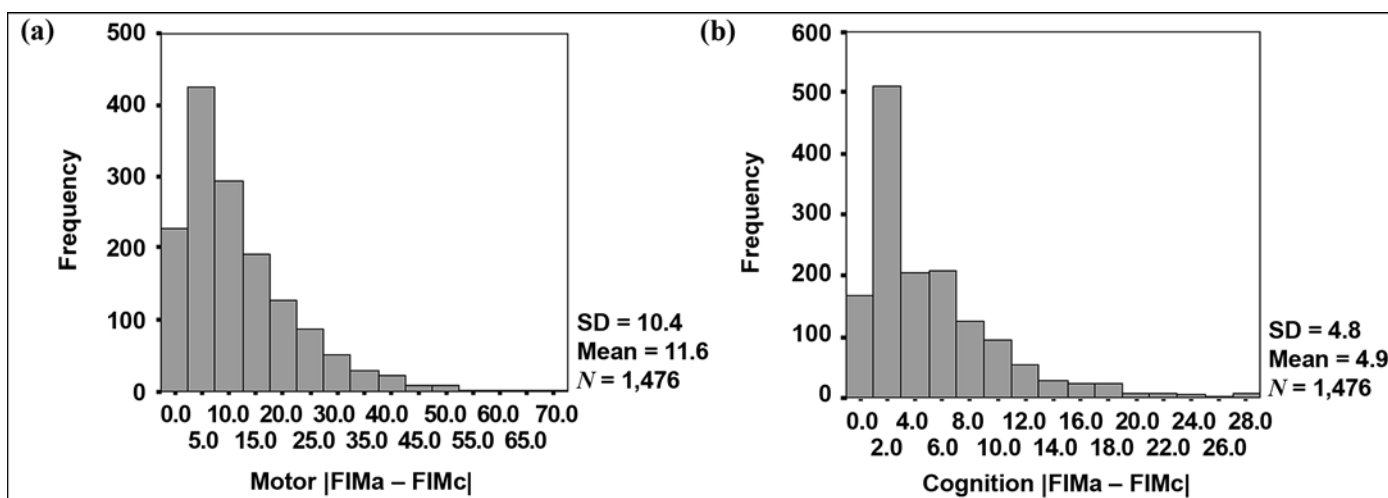
Figure 2.
Point difference between actual and converted Functional Independence Measure (FIM) score distribution. (a) |FIMa – FIMc| motor score distribution. (b) |FIMa – FIMc| cognition score distribution. FIMa = actual FIM (score), FIMc = Minimum Data Set-derived FIM (score), SD = standard deviation.

1071

WANG et al. Validation of FIM-MDS crosswalk

Validation at Classification Level

The FIMa and the FIMc were used to classify patients into different FRGs based on the same classification algorithm. **Tables 3–5** presents results from FRG classifications for individuals with stroke, amputation, and orthopedic impairment, respectively. FRGa indicates the classification results based on the FIMa scores, whereas FRGc represents the classification results based on the FIMc scores.

For individuals with stroke, the FRGs classified individuals into nine categories (**Table 3**). Chi-square statistics showed a significant association between the classification results ($\chi^2 = 1,232.6$, degrees of freedom [df] = 64, $p <$ 0.001). Kappa analysis demonstrated a fair strength of agreement ($\kappa = 0.37$). About 44.0 percent of patients were classified into the same FRGs, 67.0 percent into FRGs within ±1 FRG level, and 80.5 percent into FRGs within ±2 FRG levels.

For individuals with amputation, patients were classified into one of two FRG categories (**Table 4**). Chi-square statistics showed a significant association between the classification results ($\chi^2 = 120.6$, $df = 1$, $p < 0.001$). Kappa analysis demonstrated a substantial strength of agreement ($\kappa = 0.66$). Approximately 83.1 percent of subjects were classified into the same FRG.

Table 2.

Summary of validation results of actual FIM (FIMa) scores and MDS-derived FIM (FIMc) scores at individual patient level.

FIM Subscale	Mean ± SD	Wilcoxon*	Correlation† (r)	Point Difference |FIMa – FIMc|		
				Mean ± SD	% ≤5 points	% ≤10 points
Motor						
FIMa	55.7 ± 24.2	$z = -4.11, p < 0.001$	0.79	11.6 ± 10.4	33.7	56.9
FIMc	54.4 ± 23.5					
Cognition						
FIMa	27.0 ± 9.0	$z = -2.21, p = 0.03$	0.67	4.9 ± 4.8	67.1	87.7
FIMc	27.1 ± 7.8					

*Nonparametric Wilcoxon signed rank test.
†Correlation between FIMa and FIMc scores.
FIM = Functional Independence Measure, MDS = Minimum Data Set, SD = standard deviation.

Table 3.

Functional-related group (FRG) classification for sample of patients with stroke.

FRGc	FRGa									Total
	1	2	3	4	5	6	7	8	9	
1	88	0	0	13	4	5	1	0	2	113
2	5	65	15	3	2	3	0	0	0	93
3	1	29	39	15	9	1	4	1	0	99
4	14	6	7	32	21	15	14	7	10	126
5	6	1	3	12	13	15	11	6	3	70
6	3	4	1	4	8	6	11	4	10	51
7	3	0	1	9	8	7	51	5	20	104
8	3	0	2	3	5	7	5	7	14	46
9	0	1	1	5	3	6	17	14	55	102
Total	123	106	69	96	73	65	114	44	114	804

Note: 44.0% of patients were classified into same FRG, 67.0% one level apart, and 80.5% two levels apart; $\chi^2 = 1,232.6$, $df = 64$, $p < 0.001$; $\kappa = 0.37$.
df = degrees of freedom, FRGa = FRG based on actual Functional Independence Measure (FIM) score, FRGc = FRG based on Minimum Data Set-derived FIM score.

Table 4.
Functional-related group (FRG) classification for sample of patients with amputation.

FRGc	FRGa		Total
	1	2	
1	112	33	145
2	12	109	121
Total	124	142	266

Note: 83.1% of patients were classified into same FRG; $\chi^2 = 120.6$, $df = 1$, $p < 0.001$; $\kappa = 0.66$.
df = degrees of freedom, FRGa = FRG based on actual Functional Independence Measure (FIM) score, FRGc = FRG based on Minimum Data Set-derived FIM score.

Table 5.
Functional-related group (FRG) classification for sample of patients with orthopedic impairment.

FRGc	FRGa							Total
	1	2	3	4	5	6	7	
1	2	3	0	0	0	0	0	5
2	1	17	3	4	1	3	6	35
3	0	4	8	3	0	1	17	33
4	1	2	8	40	1	1	4	57
5	0	0	2	0	12	0	34	48
6	0	1	0	0	3	3	16	23
7	0	3	7	1	22	9	114	156
Total	4	30	28	48	39	17	191	357

Note: 55.0% of patients were classified into same FRGs, 69.2% one level apart, 87.4% two levels apart; $\chi^2 = 433.4$, $df = 36$, $p < 0.001$; $\kappa = 0.37$.
df = degrees of freedom, FRGa = FRG based on actual Functional Independence Measure (FIM) score, FRGc = FRG based on Minimum Data Set-derived FIM score.

For individuals with orthopedic impairment, seven FRG classification levels were used (**Table 5**). Chi-square statistics exhibited a significant association between the classification results ($\chi^2 = 433.4$, $df = 36$, $p < 0.001$) with a fair strength of agreement ($\kappa = 0.37$). About 55.0 percent of patients were classified into the same FRGs, 69.2 percent into FRGs within ±1 level, and 87.4 percent into FRGs within ±2 levels.

Validation at Facility Level
Lastly, we evaluated the FIM-MDS crosswalk conversion algorithm at the facility level by comparing FIMa and FIMc by facility. **Table 6** lists the validation results at the facility level. In general, the mean differences between FIMa and FIMc varied from a low of 1.41 (facility E) to a high of 11.30 (facility A). Since the distribu-

tions of the FIMa and FIMc scores also deviated from normal distributions, nonparametric statistics were used to test the equivalence of the score distributions. The results from the Wilcoxon signed rank test showed an acceptable equivalence in three out of five facilities. The correlations between FIMa and FIMc motor scores also varied from facility to facility (ranging from 0.58 to 0.89).

The cognition scale showed more consistent results. Mean differences between FIMa and FIMc cognition scores ranged from 1.51 to 3.82 points. Nonetheless, the nonparametric Wilcoxon signed rank test revealed statistical significance for all facilities ($p < 0.05$). The correlations between FIMa and FIMc cognition scores also varied from facility to facility (ranging from 0.42 to 0.80).

DISCUSSION

This study evaluated the FIM-MDS crosswalk conversion algorithm developed by Velozo and colleagues [14]. We obtained "mixed" findings from the validity testing of the FIM-MDS motor and cognition crosswalks.

Several results supported the feasibility of developing FIM-MDS crosswalks. The actual FIM and MDS scores moderately correlated at –0.80 for the motor scales and –0.66 for the cognition scales. Results showed that the mean FIMc scores were very close to the FIMa scores, with a small difference of 1.3 points deviation for the motor scale and 0.1 points for the cognition scale. From the FRG classification system, chi-square statistics showed a significant association between the FRG classification results based on FIMa and FIMc scores, with a fair (0.37) to substantial (0.66) strength of agreement. Meanwhile, at the facility level four out of five facilities had just 1- to 3-point differences between the mean FIMa and FIMc scores.

While sample distributions were similar, individual score comparisons fell short of expectations. Only 34 percent of the FIMc motor scores were within 5 points of the FIMa scores and 56.9 percent were within 10 points. For the cognition scale, 67 percent of the FIMa and FIMc cognition scores showed a 5-point difference or less and 87.7 percent show a 10-point difference or less. The results for the FIM cognition scale may be inflated because of a severe ceiling effect. The results from nonparametric procedures did not support the hypothesis that the actual and converted scores had the same score distributions

1073

WANG et al. Validation of FIM-MDS crosswalk

Table 6.
Validation results of actual FIM (FIMa) and MDS-derived FIM (FIMc) scores at facility level.

Facility	Mean ± SD	Difference[*]	Correlation[†] (r)	Wilcoxon[‡]
A (n = 63)				
FIMa Motor	77.21 ± 13.54	11.30	0.62	z = –5.64, p < 0.001[§]
FIMc Motor	65.91 ± 12.78			
FIMa Cognition	31.30 ± 5.87	2.84	0.80	z = 5.07, p < 0.001[*]
FIMc Cognition	28.46 ± 6.57			
B (n = 53)				
FIMa Motor	50.47 ± 18.91	–3.53	0.77	z = –2.14, p = 0.03[§]
FIMc Motor	54.00 ± 19.28			
FIMa Cognition	30.30 ± 7.37	1.51	0.80	z = –3.07, p = 0.002[§]
FIMc Cognition	28.79 ± 8.89			
C (n = 57)				
FIMa Motor	70.49 ± 17.42	2.08	0.70	z = –1.04, p = 0.299
FIMc Motor	68.41 ± 18.93			
FIMa Cognition	31.11 ± 6.49	2.86	0.53	z = –3.97, p < 0.001[§]
FIMc Cognition	28.25 ± 4.60			
D (n = 72)				
FIMa Motor	73.96 ± 15.61	–1.96	0.58	z = –0.54, p = 0.59
FIMc Motor	75.92 ± 15.83			
FIMa Cognition	28.90 ± 6.88	–3.82	0.42	z = –4.63, p < 0.001[§]
FIMc Cognition	32.72 ± 2.89			
E (n = 71)				
FIMa Motor	52.07 ± 23.06	1.41	0.89	z = –1.00, p = 0.31
FIMc Motor	50.66 ± 23.31			
FIMa Cognition	24.10 ± 8.97	–1.82	0.66	z = –2.01, p = 0.04[§]
FIMc Cognition	25.92 ± 7.38			

[*]Mean of |FIMa – FIMc| scores.
[†]Correlations between FIMa and FIMc scores.
[‡]Nonparametric Wilcoxon signed rank test.
[§]Indicates statistical significance with $p < 0.05$.
FIM = Functional Independence Measure, MDS = Minimum Data Set, SD = standard deviation.

for either the motor or cognition scales. Meanwhile, the validation results at the facility level varied. The correlation coefficient between the FIMa and FIMc scores varied from moderate to strong for the motor (0.58 to 0.89) and cognition scales (0.42 to 0.80). While four of the five facilities demonstrated 1- to 3-point differences between the mean FIMa and FIMc scores, one facility showed an 11.3-point difference in the motor scale.

The "mixed" findings from the validity testing of the FIM-MDS motor and cognition crosswalks leave considerable questions as to the extent to which crosswalks should be implemented. While the relatively strong correlations between the actual and converted scores support the use of a crosswalk, the significant differences found between the actual and converted score distributions suggested that the crosswalk may not have an adequate level of accuracy for decision-making in healthcare settings. The similar mean FIMa and FIMc scores suggest that the crosswalks may demonstrate "population equivalence" and may have adequate accuracy for monitoring of facility-level outcomes and for research involving large sample sizes. However, the results varied between facilities and the validation at the facility level did not completely support this conclusion. The facility-level testing was done with sample sizes just above 50 subjects. Larger sample sizes should be analyzed in future studies to investigate whether applying crosswalks at the facility level is feasible.

Our validation results based on the FRG classification system were slightly less promising than those found by Buchanan and colleagues in their 2004 study of a crosswalk between the FIM and MDS-PAC [11]. They evaluated their FIM-MDS-PAC conversion system by classifying approximately 3,200 subjects into CMGs by using a prospective study. The FIM and the MSD-PAC to FIM scales mapped 53 percent of cases into the same CMG; approximately 84 percent were classified within one CMG. In this study, about 44 percent (stroke), 83 percent (amputation), and 55 percent (orthopedic) were classified into the same FRG and 67 percent (stroke), 100 percent (amputation), and 69.2 percent (orthopedic) within one FRG level. Buchanan et al. further evaluated the payment implications of substituting the MDS-PAC for the FIM. They found the mean payment difference between these two instruments was not significantly different from zero [11]. However, the SD of differences was large ($1,960) and nearly 20 percent of the facilities had revenue shifts larger than 10 percent of the original cost. Similar cost comparisons were not made in the present study because of the unavailability of cost information. Note that the present study, as compared with Buchanan and colleagues' findings, had different methodologies and sampling (e.g., differences in FRG vs CMG calculations, as well as secondary analysis of VA data vs prospective data collection).

The effectiveness of crosswalks may depend on sources of error associated with the FIM and MDS. This sample was selected based on the criterion of "5 days or less" between administrations of the FIM and MDS. Based on the requirements that clinical staff have 3 days to complete the FIM assessment after patients are admitted or discharged from an IRF and that MDS coordinators have 14 days to finish the MDS evaluation, the "5-day" criteria used in this study may be considered stringent. This study was based on the assumption that a patient's functional status remains unchanged during this 5-day period. This assumption is unlikely to be supported in clinical practice. However, as we further filtered the data with the criterion of fewer than 5 days between FIM and MDS assessment dates, the number of records available for analysis dramatically decreased. We also performed analyses with fewer days between administrations of the assessments. Even so, the more stringent criteria did not demonstrate better validation results.

To ensure the quality of the data, clinicians must obtain 80 percent or above passing scores on the FIM mastery test to be qualified for entering FIM data into the VA national database. Moreover, clinicians are retested every 2 years with a different version of the examination to ensure their rating accuracy of the FIM. A number of studies have shown that the FIM has good interrater or test-retest reliability (>0.90) [21–24]. In contrast, for MDS assessments, each unit's assessment coordinator may obtain information from other staff orally or have specific professionals rate relevant items. The raters do not necessarily receive formal training, and no qualification examination is required. Since the MDS administration is much more extensive than the FIM, assessment burden may result in assessments being completed quickly instead of accurately, which further introduces rater error. The MDS items have been shown to have high interrater reliability with walking items (weighted $\kappa = 0.89$–0.92) [25], overall ADL items ($r = 0.99$), and cognition items ($r = 0.80$) [26]. However, few MDS studies focused on physical functioning and cognition/communication domains and information regarding the psychometrics of these two sections is insufficient. To improve the application of the FIM-MDS crosswalk conversion algorithm at the clinical sites, improvement in the quality of test administration and rater training are likely to be required.

This validation study has several limitations. First, the data sets did not have sufficient information about assessment type (e.g., admission, discharge, follow-up) and discharge reasons. For patients who transferred from an IRF to an SNF, their condition may have been different from those who transferred from an SNF to an IRF. For instance, a transfer from one facility to another might be due to a significant change in the patient's medical or functional status or merely due to the need for continued care. Second, the present validation study only focused on three major rehabilitation impairment groups. More diverse impairment groups and more representative samples should be explored in the development of FIM-MDS crosswalk conversion algorithms. Third, fewer days between assessments dates should be explored in the future to minimize the potential for functional status change during the transfer period.

The failure of the FIM-MDS crosswalks in the present study to demonstrate "individual equivalence" (i.e., relatively low percentage of actual and converted scores being within 5 points of each other), suggests that the crosswalks do not have adequate accuracy to monitor individual patients who transfer from facilities that use the FIM (e.g., IRFs) or from facilities that use the MDS (e.g., SNFs). While these validation results fail to support

1075

WANG et al. Validation of FIM-MDS crosswalk

the use of a crosswalk to monitor individuals across the continuum of care, the variance might be as much a function of the error in global measures per se as it is a function of error in the crosswalk. That is, variance in use of the same measure (e.g., only the FIM or only the MDS/MDS-PAC) by different raters and across different facilities could result in different FRG/CMG classifications and revenue shift differences similar to those found by Buchanan and colleagues in their study of the FIM-MDS-PAC crosswalk. That is, while using a single comprehensive outcome assessment instrument for postacute care may appear to be the best mechanism for monitoring patient outcomes across postacute settings, inherent error in these measures may lead to the type of inconsistencies found with the crosswalks. A direct comparison of a single measure to crosswalk conversions may be necessary to determine the best method for monitoring outcomes across the continuum of care. The effectiveness of a single measure or of crosswalk conversions may ultimately depend on the quality of the data. Effective monitoring of patients across the continuum of care, whether via a single measure or crosswalks, may ultimately depend on more rigorous standardization and more extensive training of clinical raters.

CONCLUSIONS

To evaluate and track changes of functional status within the course of rehabilitative healthcare services, researchers and clinicians must integrate information across the continuum of care. The crosswalk algorithm demonstrated a convenient way to achieve score comparisons across different rehabilitation settings. However, the effectiveness of a single measure or of crosswalk conversions may ultimately depend on the quality of the data.

ACKNOWLEDGMENTS

REFERENCES

1. Ottenbacher KJ, Hsu Y, Granger CV, Fiedler RC. The reliability of the Functional Independence Measure: A quantitative review. Arch Phys Med Rehabil. 1996;77(12):1226–32. [PMID: 8976303]

2. Dahl TH. International classification of functioning, disability and health: An introduction and discussion of its potential impact on rehabilitation services and research. J Rehabil Med. 2002;34(5):201–4. [PMID: 12392233]

3. Granger CV, Hamilton BB. The Functional Independence Measure. In: McDowell I, Newell C, editors. Measuring health: A guide to rating scales and questionnaires. 2nd ed. New York (NY): Oxford University Press; 1987. p. 115–21.

4. Carter GM, Relles DA, Ridgeway GK, Rimes CM. Measuring function for Medicare inpatient rehabilitation payment. Health Care Financ Rev. 2003;24(3):25–44. [PMID: 12894633]

5. Rehab Management [Internet]. Los Angeles (CA): Ascent Media; c2008 [updated 2001 Mar; cited 2006 May]. Murer CG. Trends and issues; [about 4 screens]. Available from: http://www.rehabpub.com/departments/32001/5.asp/.

6. Morris JN, Hawes C, Fries BE, Phillips CD, Mor V, Katz S, Murphy K, Drugovich ML, Friedlob AS. Designing the national resident assessment instrument for nursing homes. Gerontologist. 1990;30(3):293–307. [PMID: 2354790]

7. Haley SM, Coster WJ, Andres PL, Ludlow LH, Ni P, Bond TL, Sinclair SJ, Jette AM. Activity outcome measurement for postacute care. Med Care. 2004;42(1 Suppl):I49–61. [PMID: 14707755]

8. Fries BE, Schneider DP, Foley WJ, Gavazzi M, Burke R, Cornelius E. Refining a case-mix measure for nursing homes: Resource Utilization Groups (RUG-III). Med Care. 1994;32(7):668–85. [PMID: 8028403]

9. U.S. Centers for Medicare and Medicaid Services. Revised Long Term Care Resident Assessment Instrument user's manual: Version 2.0. Washington (DC): American Health Care Association; 2002.

10. Rogers JC, Gwinn SM, Holm MB. Comparing activities of daily living assessment instruments: FIM, MDS, OASIS, MDS-PAC. Phys Occup Ther Geriatr. 2001;18(3):1–26.

11. Buchanan JL, Andres PL, Haley SM, Paddock SM, Zaslavsky AM. Evaluating the planned substitution of the Minimum Data Set-Post Acute Care for use in the rehabilitation hospital prospective payment system. Med Care. 2004;42(2):155–63. [PMID: 14734953]

12. McMullan M. Regulatory comment letters. Re: HCFA-1069-P Medicare prospective payment system for inpatient rehabilitation facilities proposed rule (65 Federal Register 66304) [Internet]. Chicago (IL): American Hospital Association; c2000 [cited 2006 Oct 15]. Available from: http://www.aha.org/aha/letter/2001/010126-cl-65fr66304.html/.

13. Grabois M. Open letter from ACRM to HCFA on proposed Medicare PPS. American Congress of Rehabilitation Medicine. Health Care Financing Administration. Arch Phys Med Rehabil. 2001;82(4):567–69. [PMID: 11310452]

14. Velozo CA, Byers KL, Wang YC, Joseph BR. Translating measures across the continuum of care: Using Rasch analysis to create a crosswalk between the Functional Independence Measure and the Minimum Data Set. J Rehabil Res Dev. 2007;44(3):467–78. [PMID: 18247243]

15. Kolen MJ, Brennan RL. Test equating, scaling, and linking: Methods and practices. New York (NY): Springer; 2004.

16. Buchanan JL, Andres PL, Haley SM, Paddock SM, Zaslavsky AM. An assessment tool translation study. Health Care Financ Rev. 2003;24(3):45–60. [PMID: 12894634]

17. Jette AM, Haley SM, Ni PS. Comparison of functional status tools used in post-acute care. Health Care Financ Rev. 2003;24(3):13–24. [PMID: 12894632]

18. Stineman MG, Escarce JJ, Goin JE, Hamilton BB, Granger CV, Williams SV. A case-mix classification system for medical rehabilitation. Med Care. 1994;32(4):366–79. [PMID: 8139301]

19. Stineman MG. Case-mix measurement in medical rehabilitation. Arch Phys Med Rehabil. 1995;76:1163–70. [PMID: 8540795]

20. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33(1):159–74. [PMID: 843571]

21. Ottenbacher KJ, Mann WC, Granger CV, Tomita M, Hurren D, Charvat B. Inter-rater agreement and stability of functional assessment in the community-based elderly. Arch Phys Med Rehabil. 1994;75(12):1297–1301. [PMID: 7993167]

22. Hamilton BB, Laughlin JA, Fiedler RC, Granger CV. Inter-rater reliability of the 7-level functional independence measure (FIM). Scand J Rehabil Med. 1994;26(3):115–19. [PMID: 7801060]

23. Ottenbacher KJ, Taylor ET, Msall ME, Braun S, Lane SJ, Granger CV, Lyons N, Duffy LC. The stability and equivalence reliability of the Functional Independence Measure for children (WeeFIM). Dev Med Child Neurol. 1996;38(10):907–16. [PMID: 8870612]

24. Kidd D, Stewart G, Baldry J, Johnson J, Rossiter D, Petruckevitch A, Thompson AJ. The Functional Independence Measure: A comparative validity and reliability study. Disabil Rehabil. 1995;17(1):10–14. [PMID: 7858276]

25. Morris JN, Nonemaker S, Murphy K, Hawes C, Fries BE, Mor V, Phillips C. A commitment to change: Revision of HCFA's RAI. J Am Geriatr Soc. 1997;45(8):1011–16. [PMID: 9256856]

26. Casten R, Lawton MP, Parmelee PA, Kleban MH. Psychometric characteristics of the Minimum Data Set I: Confirmatory factor analysis. J Am Geriatr Soc. 1998;46(6):726–35. [PMID: 9625189]